Use of Alternate Assessment Results in Reporting and Accountability Systems: Conditions for Use Based on Research and Practice

NCEO Synthesis Report 43

Published by the National Center on Educational Outcomes

Prepared by:

Rachel Quenemoen • Susan Rigney • Martha Thurlow

May 2002

Executive Summary

State assessment systems must address both technical issues and policy issues as assessments and accountability practices are developed and implemented. These technical and policy issues have been expanded from traditional large-scale assessment to the new alternate assessment approaches required by law and developed in every state. The primary purpose of federally required alternate assessments in state or district assessment systems is the same as the primary purpose of federally required large-scale assessments, that is, accountability. The purpose of both is to provide valid and reliable assessment data that accurately reflect the state's learning standards, and that indicate how a school, district, or state is doing in terms of overall student performance. From that information, schools can make broad policy decisions that improve schooling practices so all students are successful.

As these approaches to assessment are implemented, states have raised questions about how results from disparate tests can be combined for the primary purpose of accountability. This report reviews our current understanding of the technical and policy considerations involved in high quality alternate assessment. Based on research from early implementation and what are considered to be best practice approaches, this synthesis describes five steps in alternate assessment test development processes that allow interpretation and use of results in reporting and accountability.

Responsible use of any assessment includes documentation of its purpose and technical quality. In the development of alternate assessments, many states have worked with researchers and practitioners to assure that the performance measures used with students who have significant disabilities are also defensible in terms of reliability and validity. *Once these measurement issues have been addressed, the issue of how to include the results in school accountability decisions becomes primarily a policy decision.* There are several different ways that can occur. Three state examples illustrate different methods of incorporating alternate assessment results in accountability systems.

In general, these conclusions for practice can be asserted:

1. Choosing and implementing the best option for each state requires collaboration among state offices for policy decisions, funding, and teacher training.
2. Careful and thoughtful delineation of the individualized nature of IEP team planning along with the system measurement and reporting requirements must occur.
3. Understanding policy requirements also requires understanding of technical issues.
4. Increased focus on credibility of alternate assessment results is required, including documentation of the reliability and validity of a state's alternate assessments, and increased rigor of standard setting processes for alternate assessment.
5. It is essential to consider how a state's approach reflects the assumption of standards-based reform that all children can learn, and schools can be held accountable for their learning.

I was pretty worried about alternate assessment requirements to begin with, but as we went through training, several

of us started talking about how really alternate assessment just asked us to do what we love and do best – teach children so that they learn, and can show what they know and can do! I know I have to do my job well because my students are going to be counted, and I am accountable for what they have learned. It focused my instruction, and I have more clarity on where I need to head with each student. And now I have data to talk about. I don't just say 'She looks better.' I have something that is measurable. It's hard work including students who have never been part of the system. It took a lot of advocacy, which has really reenergized my work.

(Thompson, Quenemoen, Thurlow, & Ysseldyke, 2001, composite of teacher comments from surveys after first year of implementation).

Overview

During the past decade, the emergence of standards-based reform has generated debate among policymakers and educators about how to ensure that all students benefit from the reform. Advocates of a standards-based system assume that all students can be expected to attain high standards of learning, although some may need more time and varied instruction. As a result of federal legislation (IDEA and Title I), state-level policymakers and educators have been required to determine not whether, but how, all students will participate fully in instruction, assessment and school accountability based on high standards. Educators have had to translate the policy commitment of high expectations for all students into practice by:

1. Defining content standards describing what all students should know and be able to do, and defining acceptable levels of performance.
2. Ensuring that all students have the opportunities to learn this content.
3. Developing technically sound assessments to measure student performance.
4. Developing methods of using the assessment results to hold schools accountable for students' learning.

State assessment systems must address both technical issues and policy issues as assessments and accountability practices are developed and implemented. In a standards-based system, the assessments that are used to hold schools accountable must accurately reflect the state's learning standards in order to be valid. The traditional technical issues of test validity and reliability have been broadened to include consideration of the impact of testing, often referred to as consequential validity (Brualdi, 1999; Messick, 1989).

These technical and policy issues have been expanded from traditional large-scale assessment to the new alternate assessment approaches required by law and developed in every state. Alternate assessments provide a mechanism for students with the most significant disabilities to be included in the assessment system. The primary purpose of federally required alternate assessments in state or district assessment systems is the same as the primary purpose of federally required large-scale assessments, that is, accountability. The purpose of both is to provide valid and reliable assessment data that accurately reflect the state's learning standards, and that indicate how a school, district, or state is doing in terms of overall student performance. From that information, schools can make broad policy decisions that improve schooling practices so all students are successful.

As these approaches to assessment are implemented, states have raised questions about how results from disparate tests can be combined for the primary purpose of accountability. This report reviews our current understanding of the technical and policy considerations involved in high quality alternate assessment. Based on research from early implementation and what are considered to be best practice approaches, this synthesis describes alternate assessment test development processes that allow interpretation and use of results in reporting and accountability; examples of state solutions are presented, and recommendations for practice are provided.

Alternate Assessment: A Rethinking of Traditional Test Development Processes

Alternate assessments are the focus of controversy for a number of reasons. The nature of the controversy surrounding alternate assessments reflects, in part, the nature of the alternate assessments that are currently in use, and the characteristics of the students for whom they are designed. To really understand alternate assessments, it is important to think more broadly about the inclusion of students with disabilities in educational assessments over the past century. Not only has there been limited inclusion of most students with disabilities, but those students with the most significant disabilities have been excluded without exception. We have had at least a half a century to fine-tune how to assess "average" students, but only a few years to devote to a similar development process for students with complex disabilities. As our country refined its measures of educational attainment to focus on standards and what "proficiency" means for the general population of students, research-based efforts increased, resulting in a literature base that clarifies what proficient performance is and what instructional strategies and organizational features are necessary for ensuring that students reach proficiency. (See Mid-Continent Regional Laboratory Web site for a comprehensive bibliography of this literature base: http://www.mcrel.org/standards-benchmarks/docs/reference.asp)

This process is just beginning for alternate assessments designed for students with significant disabilities. For the most part, but not entirely, the analogous development process is being carried out by states. Research-based efforts (see Browder, 2001; Kleinert & Kearns, 2001; Thompson, Quenemoen, Thurlow, & Ysseldyke, 2001) are clarifying the notion of proficiency for students with significant disabilities. Some states are beginning to produce research-based evidence where quality curriculum outcomes are definable and measurable for students with significant disabilities. Just as traditional assessment processes and

definitions of proficiency have implicit assumptions (e.g., a proficient student will perform well in the next grade or educational level), so do alternate assessment processes. These assumptions have grown out of our professional historical understanding of what proficiency looks like for these students.

History and Development of Alternate Assessments

Because alternate assessments are new, it is important to revisit how they have emerged within a historical context. The definition of optimal outcomes for students with significant disabilities has shifted since 1975 when these students were first guaranteed a free appropriate public education. In the 1980s there was a move away from a notion of "developmental sequence," where all students were expected to progress through a "normal" developmental sequence of infanthood, toddlerhood, and so forth. Instead, there emerged a focus on functional domains that were necessary for success in current and future environments. As a field, we stopped teaching students at their "mental age" (e.g., two year old developmental level), and focused on chronological age functions, preparing them for practical application of school based learning. During that time, however, we seemed to have lost touch with academic goals and instruction – focusing on community-based activities instead of concrete learning of skills and knowledge in a variety of settings.

With the advent of "inclusion" models in the 1990s, there was a tendency to focus on the importance of "social" progress from contact with peers; this was often valued over numeracy or literacy skills in a variety of settings. In many cases academics was addressed only incidentally. Still, some students remained isolated in self-contained classrooms, being kept warm and comfortable by well intentioned, caring aides, but with very limited "learning" taking place. (See Browder, 2001, pages 13-17, for an excellent summary of this historical sequence, and specific linkages to how large-scale assessments for students with significant disabilities are being developed.)

The development of high quality alternate assessments required a reexamination of these sometimes competing approaches, and put pressure on states and schools to articulate precisely what "learning" meant for this population. As a result, and with thoughtful commitment to bringing all students into the opportunities of standards-based reform, the special education and assessment communities have identified a set of steps for the development of alternate assessments that are analogous to the process used in developing general assessments (see Table 1). In states where a thoughtful process occurred, the five steps are evident.

Table 1. Steps in Development of Alternate Assessments

Careful stakeholder and policymaker development of desired student outcomes for the population, reflecting the best understanding of research and practice.
Careful development, testing, and refinement of assessment methods.
Scoring of evidence according to professionally accepted standards.
Standard-setting process to allow use of results in reporting and accountability systems.
Continuous improvement of the assessment process.

1. Careful stakeholder and policymaker development of desired student outcomes for the population, reflecting the best understanding of research and practice. These articulated student outcomes are essential to the development of a rigorous assessment system. It is in this step where thoughtful linkages to state standards must be articulated, and it is in this step that many states have extended their content standards to ensure that students with significant cognitive disabilities have access to, and make progress in, the general curriculum. These desired student outcomes, linked to the content standards defined for all students, are reflected in the rubrics or scoring criteria used to score the evidence, regardless of the type of evidence gathered. The development of draft rubrics/criteria and subsequent refinements is how these outcomes are operationalized in the measurement process. The desired student outcomes are also reflected in the achievement descriptors and ultimately in the achievement levels. Although there is general professional congruence on what these desired outcomes may be (Browder, 2001; Kleinert & Kearns, 1999), the precise desired outcomes vary from state to state, just as do the content standards, rubrics, descriptors, and achievement levels for the general assessment. Research is underway to document the variation in articulated outcomes and rubrics or scoring criteria across states.

2. Careful development, testing, and refinement of assessment methods. Typically, the assessment methods are ways to gather evidence, resulting in a portfolio process, assessment instruments, or other approach that will yield high quality evidence. This has been the focus of many states' efforts over the past few years. We have seen shifts in methodology after pilot years or first years of implementation (Thompson & Thurlow, 2001). Just as for the general assessment, the development of a rigorous, valid, and reliable instrument takes commitment and time, but is one step in the process, not the only step. Not all states are at this point. In states with a complete alternate assessment development process, extensive training and support for teachers, parents, and other IEP team members also occurs.

3. Scoring of evidence according to professionally accepted standards. Once assessment evidence is gathered, thoughtful states have engaged in rigorous scoring procedures following rigorous professional standards for assessment scoring. Scoring training is provided, scorers demonstrate their competency, and inter-rater reliability tests and rechecks of scorer competency occur throughout the process. Dual scoring, third party tie breakers - all tools of the assessment trade - should be in evidence in this step (AERA/APA/NCME, 1999; Thompson, Quenemoen, Thurlow, & Ysseldyke, 2001).

4. Standard-setting process to allow use of results in reporting and accountability systems. Once scores are assigned, states with a complete alternate assessment development process have moved on to typical steps in standard-setting processes, such as reviewing student work to identify initial "bands" of possible cut scores across achievement descriptors, followed by panel reviews of student work to arrive at cut scores or panel reviews of hypothetical cut score parameters (Cizek, 2001; Roeber, 2002; Thompson, Quenemoen, Thurlow, & Ysseldyke, 2001).

5. Continuous improvement of the assessment process. In every state that has accomplished this full process, we have seen revision of rubrics, editing of achievement descriptors, and increased focus on training of teachers and other IEP members prior to second year implementation. And in states where there has been an emerging body of research, we have seen promising evidence of reliability, validity, and more importantly, a direct link between improving alternate assessment scores and improvements in instruction (Kamfer, Horvath, Kleinert, & Kearns, 2001; Kleinert, Kennedy, & Kearns, 1999; Quenemoen, Massanari, Thompson, & Thurlow, 2000; Turner, Baldwin, Kleinert, & Kearns, 2000).

This type of development process for alternate assessments is comparable to that used for regular assessments. A growing body of literature is demonstrating an emerging consensus on these characteristics of good alternate assessments (see Appendix).

Several states have followed this cycle of "test development" with their alternate assessments. Kentucky is the leader, in part because it was the first state to develop an alternate assessment. It has the broadest base of research to back its progress. There are other states as well. Massachusetts, Arkansas, and West Virginia discussed their approaches at a recent meeting of state alternate assessment staff (ASES SCASS), to name just some of the states that have completed a full assessment development cycle that includes standard setting. These states are in good position to build on this rigorous and research-based approach to improve outcomes for students with the most significant disabilities.

Status of State Alternate Assessments

NCEO has documented the development of alternate assessments in states for several years (Thompson, Erickson, Thurlow, Ysseldyke, & Callender, 1999: Thompson & Thurlow, 2000, 2001). In some states, the alternate assessment is essentially a paper and pencil test adjusted to a lower level of standards. It is evident in these states that the alternate assessment is intended for a group of students that has not kept up with the majority of students of the same age, either because of low cognitive functioning or because of other disabilities. They are students for whom the same general instructional goals have been defined; they are just meeting them at slower rates than other students. In some states, the alternate assessment is simply a teacher checklist of developmental skills. The group of students for whom these types of assessments are intended vary with the nature of the checklist, which may focus on adaptive behavior in some states and on reading skills in others.

In most states, the alternate assessment consists of a body of evidence collected by educators, parents, and the student to demonstrate and document the student's skills and growth toward state standards; sometimes these alternate assessments also incorporate characteristics of educational supports that the student receives. In most of the states with this type of alternate assessment, it is clear that the students for whom the alternate assessment is intended have very complex disabilities – most often significant cognitive disabilities.

Current Challenges in Alternate Assessments

The variability in alternate assessments makes them more difficult to understand. Still, we know that alternate assessments are a problem when they are used for a broader group of students than they should be, when they are used to lower standards, and when they are a way to exclude students from the accountability system. While it is difficult to determine exactly how many students should be in an alternate assessment, the finding that states are planning on having anywhere from 5% to more than 40% of their students with disabilities in their alternate assessment system (Thompson & Thurlow, 1999) suggests that there is a problem with defining a common target group for the assessment.

Alternate assessments should not be used to lower standards for students with disabilities. NCEO's analysis of state's alternate assessments (Thompson & Thurlow, 2001) revealed an increasing number of states with a two-prong alternate assessment – one prong for students with significant and complex disabilities, and the other for students not functioning on grade level. Not only does this approach increase the number of students in the alternate assessment, but it brings in students who indeed can take the regular assessment, though they may not perform very well on it. An alternate assessment should not be for those students not expected to do well. It should be those who are working toward the essence of standards, where standards have been viewed broadly to encompass those students with very complex disabilities. To the extent that states develop clear guidelines for who should participate in the alternate assessment, and to the extent that those guidelines define a group of students with significant, complex disabilities, then it is possible to hold alternate assessment students to high standards and to document how they can reach proficient status. By doing this, it is then possible to include these students in the accountability

system in a way that values their attainment of expanded standards just as much as the system values the attainment of students in the regular assessment.

Alternate assessments should not be a way to exclude students from the accountability system. If the alternate assessment is an avenue to exclusion from accountability, we are likely to see more and more students inappropriately pushed into the alternate assessment simply because they are not expected to perform well. Making sure that all students are in the accountability system – that all students count – is critical to avoiding corruption in accountability.

Issues Underlying Current Struggles with Including Students with Disabilities in Accountability Systems

There are many reasons why states and districts have resisted the participation of students with disabilities in assessments and accountability systems. Low expectations are a primary reason – they permeate education and translate into fears about emotional trauma and other types of emotional abuse from testing. Another reason is that special education's history of separating and taking care of students with disabilities has reinforced the notion that general educators do not know how to provide the instruction that these students need, and a perception that if they do not have the needed skills, then certainly they should not be held accountable.

High stakes assessments that carry significant consequences for students are a major cause of the struggle over the participation of students with disabilities in assessments and accountability systems (Heubert & Hauser, 1999). It seems reasonable to argue that students with disabilities have not had the same access as have other students to standards-based content, the general curriculum, and good instruction. There is often a desire to remove them from the accountability system – or to give them waivers – so that they can graduate or progress from one grade to the next without having to demonstrate the same knowledge and skills as other students. This may also limit system incentives to provide the same access to students with disabilities. Thus, high quality alternate assessment systems are developed within policy systems that provide the system incentives to provide the same access to all children, and these policy decisions drive the way the assessment system is designed, and how use of results in reporting and accountability is defined.

Policy Decisions: Three Different Methods of Including Results from an Alternate Assessment in the State Accountability System

Responsible use of any assessment includes documentation of its purpose and technical quality. In the development of alternate assessments, many states have worked with researchers and practitioners to assure that the performance measures used with students who have significant disabilities are also defensible in terms of reliability and validity. *Once these measurement issues have been addressed, the issue of how to include the results in school accountability decisions becomes primarily a policy decision.* There are several different ways that can occur. The following state examples reflect three of these possible methods.

The three states discussed here share a firm policy commitment to include all students with disabilities in assessment and accountability. At the policy level, these states began from the same place: all meant all, no exceptions. Each state developed an approach to alternate assessment that involves a portfolio or body of evidence. Each state assesses student performance linked to the state standards that apply to all students, with extensions or linkages. Each state is able to show reliability and validity of alternate assessment scores. And each of these states includes results from the alternate assessment in the school accountability system. Each state bases school accountability on gains in student achievement as demonstrated on state assessments over time

Kentucky and North Carolina have been prominent leaders in school reform, each with a long and productive history of thoughtful policy and rigorous technical approaches. Wyoming has adopted procedures to deal with the statistical constraints of small numbers. Each state employs a different method of incorporating results from the alternate assessment in the school accountability system. The following are very general descriptions of the approaches, presented in order to highlight key similarities and differences, and to show how the solution meets the policy requirements of holding schools accountable for high expectations for all students, including those students whose disabilities require an alternate assessment.

These descriptions apply to state accountability system designs from online descriptions or written materials publicly shared as of end of year 2001, and do not reflect changes after that date. Inclusion of these states here does not suggest their accountability approach will be approved by Title I review – each of these states has features that may or may not fully conform to the reauthorized ESEA. They all have attempted to begin with an assumption that all students count, and they have worked to build an accountability approach that ensures that all students can benefit from the required improvements resulting from accountability.

WYOMING

Wyoming assessment results are expressed as performance levels and descriptors. In order to avoid confusion or simplistic comparisons between the general and alternate assessments, results from the general assessment are expressed as four performance levels (advanced, proficient, partially proficient, novice), those from the alternate assessment are expressed as

three different performance levels (skilled, partially skilled, beginning) with different descriptors. However, they use both in a complex formula for accountability (Marion, 2001).

School accountability is a two-stage process. In the first stage, the state computes an index based on combined general assessment components and evaluates the overall gain from year to year against a long-range target. School classification is constrained by the level of participation, for example, no school can be Satisfactory with less than 98% of all students participating in the assessment system. For schools that are not making adequate gains, additional evidence is considered, including: participation rates in the alternate assessment, results in the alternate assessment; results of assessments given at grades 1 and 2, progress of students receiving Title I services, and reduction in percent of students at the "novice" level. A school can gain or lose points based on a combination of participation and progress on the alternate assessment.

- Schools receive 4 points if they made progress in the average alternate assessment score *or* they had 100% participation in the assessment system.
- Schools receive 2 points if they had no progress in average alternate assessment score or they have a 99% participation rate.
- Schools receive 0 points if they have a decline in the average alternate assessment scores and they had 98% participation.
- Schools are penalized by 2 points subtracted if they have only 97% participation.
- Schools are penalized by 4 points subtracted if they have less than 97% participation.
- Participation rule: If participation is under 95%, the school goes into school improvement regardless of other factors.

Inclusive Advantages: Wyoming's emphasis on getting all students into the system has resulted in a 99% participation rate. Schools can improve their status through improved alternate assessment scores and through full participation.

Disadvantages: Wyoming's two stage approach effectively weights alternate assessment results less than general assessment scores; for schools making adequate gains, alternate assessment results have minimal or no effect.

NORTH CAROLINA

North Carolina employs a different but also weighted approach to incorporate results from the alternate assessment in the accountability index. For the general assessment, students are tested in multiple content areas, and results from each student in each of these areas are combined in the accountability index. The performance levels are I-IV for the general assessments. The second academic assessment, the alternate assessment academic inventory (NCAAAI) for students who may not be able to take the general assessment but who are not eligible for the alternate assessment portfolio (NCAAP), has two (low and high) levels within each level - novice, apprentice, proficient, distinguished - in multiple content areas (reading skills, mathematics skills, writing skills). This academic inventory was pilot tested in 2000-01 and is scheduled for inclusion in the performance composite (only) in 2001-02. Schools were held accountable for the performance of students taking the alternate assessment portfolio by its inclusion in the performance composite in 2000-01. The levels for the portfolio are novice, apprentice, proficient, and distinguished. Students receive a score in each of four domains (Communication, Personal and Home Management, Career and Vocational, and Community); each domain score contributes one-fourth to the performance composite, so that a student's overall performance on the portfolio counts once, rather than four times in the performance composite.

The North Carolina formula is similar to Wyoming in that it has a two-stage process. In the North Carolina ABCs, there are two types of composite scores: growth and the performance composite. Alternate assessment portfolio scores are not included in the growth composite at this time; they are included in the performance composite. The total weighted growth composite for a school is the sum of the weighted growth components. Components of the model in 2000-01 included:

- EOG Reading and Math (Grades 3-8),
- 10 EOC tests using prediction formulas (Algebra I, Algebra II, Biology, Chemistry, ELPS, English I, Geometry, Physical Science, Physics, and US History),
- English II,
- College University Prep/College Tech Prep,
- Competency Test (percent change/gain from grade 8 to grade 10),
- Comprehensive Test in Reading and Mathematics (growth from grade 8 to grade 10)
- Change in ABCs dropout rate (1998-99 minus 1999-2000).

For computations of the performance composite, the total number of scores at or above achievement Level III in each subject included in the ABCs model is divided by the total number of eligible test-takers (i.e., valid scores, absent students, etc.) for all tests. Components included in 2000-01 were:

- EOG Reading and Math (Grades 3-8),
- 10 EOC tests (Algebra I, Algebra II, Biology, Chemistry, ELPS, English I, Geometry, Physical Science, Physics, and US History),
- English II,
- Comprehensive Test in Reading and Mathematics (growth form grade 8 to grade 10),

- NCAAP (grades 3-8 and 10),
- Writing (grades 4 and 7)
- Computer Skills test at grade 8.

North Carolina averages test results to arrive at the school index. Students who take the regular assessment contribute more scores to the school average than students who take the alternate assessment who are counted just once Nevertheless, all students can be included in the determination of school accountability at the second tier with one score instead of several.

Inclusive Advantages: North Carolina's approach includes all students in assessment and reporting.

Disadvantages: North Carolina's approach effectively weights alternate assessment results for students with the most significant disabilities less than scores from the general assessment. This occurs through the combination of its two-stage approach and its use of one score rather than multiple content area scores. This approach weights the scores of students with disabilities proportional to their presence in the assessment population with each assessment counting once.

KENTUCKY

Kentucky was a pioneer in the development of fully inclusive assessment and accountability systems. Kentucky has almost 100% of their students participating in the assessment system. Approximately 0.5% of the total enrollment (at the benchmark grades) participates in the alternate assessment. Working in close collaboration with University of Kentucky researchers, the state developed an alternate portfolio system aligned with the Kentucky content standards intended for all students. Their assessment system is a model for thoughtful and technically rigorous assessment.

Kentucky uses four performance levels to describe student work within a content area: novice, apprentice, proficient, and distinguished. For each student, points are awarded on the basis of the performance level attained. Results for students who do not participate in an assessment are counted as zero and averaged with all other students. Kentucky took the position that students who demonstrate "proficient" performance within the structure of the alternate assessment should make the same contribution to the school accountability index as students scoring at the proficient level on the regular assessment, so results from the alternate assessment are reported using the same terminology and point values employed in the regular assessment. In addition, because a student taking the alternate assessment is required to demonstrate achievement within multiple content areas, the overall score from the alternate is entered for each content area represented in the regular assessment. Essentially, the Kentucky accountability system can be thought of as averaging students, rather than test scores, and each student receives equal weight as the school index is computed.

Inclusive advantages: Kentucky's system has resulted in nearly 100% participation and 100% inclusion of results in reporting and accountability.

Disadvantages: There is a perception that their procedures obscure meaningful differences in performance between the regular and alternate assessments.

A Fourth but Controversial Method in Use in Some States

There is another approach used in a few states that requires general discussion. A few states set arbitrary values for all alternate assessment results: for example, the performance levels for all other students are set at 3 or 4 defined performance levels, but all alternate assessment results are restricted to the lowest of levels, or arbitrarily reported as a "0" regardless of student performance. The argument is that in order to be eligible for alternate assessment, students are performing at very low levels, thus regardless of performance, they are by definition in the lowest level. This approach might be viewed as the easiest to defend, but it is likely to defeat the purpose of school accountability for these students. Automatic lowest level scores do not yield information helpful in identifying improvements in performance in a way that holds educators responsible for their improvement.

Does it Matter How the State Incorporates the Alternate Assessment Results into the Accountability Mix?

Common objection 1: "Those students will bring our school index down."

Common objection 2: "Awarding the same number of points for successful performance on the alternate devalues proficient performance on the regular assessment."

In his examination of current state practices, Richard Hill of the National Center for the Improvement of Educational Assessment (Hill, 2001) examines approaches for including alternate assessment results in an overall index and draws some conclusions that address both of the common objections stated above. Based on accountability simulations using actual state data, Hill reached the following conclusions:

1. The impact of including scores from alternate assessment on school gains is trivial if the numbers of alternate assessment

participants remain fairly constant at the school level from year to year.
2. Making gains on the alternate comparable to gains on the regular assessment introduces little additional measurement error.
3. Including alternate assessment results in accountability appears to lead to better outcomes for the students who participate; and, if you are looking at ***gains***, the school should be entitled to equivalent reward or positive consequences for gains in alternate assessment results.

Hill identifies and compares two primary approaches to scaling results of alternate assessment for inclusion in accountability systems:

Option one: Scale results of the alternate assessment so that the value awarded for performance levels on the alternate are the same or similar to the value awarded for performance levels on the general assessment (similar to Kentucky). In states that use this procedure, improvement of results on the alternate over time is rewarded as substantially as improvement on the regular assessment. He sees this approach as more fair because it boosts the motivation to include all students, and results in gain scores that are accurate, unless you have a huge change in the number participating in the alternate from year to year in a given school.

Option two: Scale results of the assessments so that the performance levels on the alternate are at the lower end of the scale and performance levels on the regular assessment occupy the upper end of the scale, (with or without possibility of overlap between the upper performance levels on the alternate and lower levels of the regular assessment). In states that use this approach, the alternate assessment is assigned a defined value; that is, some states may scale the assessment system so that anything less than 50% is at the lowest performance level. That correctly measures output, gains are accurate – but you really do not need an alternate assessment at all with this approach, since by definition students will be performing under 50%, and thus would automatically be at on the lowest level.

Hill suggests that as states develop their accountability formula, the essential questions to be addressed are: What is fair? What will encourage the greatest improvement for every student? What seems reasonable? Given the limited evidence that the inclusion of results from alternate assessment does technical “damage” to school scores, and given the assumption that the gains for these students are important in improving outcomes for these students, Hill concludes that the decision about how the state will incorporate results from the alternate assessment into a school accountability formula is primarily a policy decision, not a technical decision.

What Do These State Practices Suggest to Policymakers?

How policy decisions are implemented in accountability calculations can vary, as seen by the three featured states, Wyoming, North Carolina, and Kentucky. All have made a policy decision to include all students in accountability; each has selected a different technical approach to putting that policy into practice. On the technical level several different approaches are workable, albeit with varying advantages and disadvantages. Based on initial work by Richard Hill, it appears these approaches meet basic technical requirements. Each of these states has also grappled with developing a policy approach that includes all students in accountability, although only Kentucky has equal weighting of all students.

There are many states that do not as yet have all students included in accountability formulas. A few states have not completed development of a technically adequate alternate assessment; those states will have to address these deficiencies immediately. But many states have alternate assessments that are aligned to state standards and that meet current technical requirements. In these states, they should continue building the technical soundness over time, but immediately work on policies that make use of the alternate assessment results in reporting and accountability. Beyond compliance issues, if alternate assessment scores are not included in accountability systems, a troubling policy message is sent that “some” students will not “count” in a reform based on the belief that all students ***can learn*****.**

Conclusions

As states decide how to integrate all scores into the accountability index, they must address core policy issues. Some questions that must be answered include:

- Have we developed assessments and accountability systems that reflect a priority on closing the achievement gap? Both Title I and IDEA target the children on the low side of the achievement gap.
- Have we thoughtfully developed an alternate assessment that is reliable and valid; and is it clearly “raising the bar?” Alternatively, in our approach have we just documented status quo of current programs, and thus cannot really address “progress?”

Each of our example states began with a firm policy commitment to include all students in assessment and accountability. Each has committed resources to development of a valid and reliable alternate assessment aligned to the same challenging content standards set for all students.

In general, these conclusions for practice can asserted:

1. Choosing and implementing the best option for each state requires collaboration among state offices for policy decisions, funding, and teacher training.
2. Careful and thoughtful delineation of the individualized nature of IEP team planning along with the system measurement and reporting requirements must occur.
3. Understanding policy requirements also requires understanding of technical issues.
4. Increased focus on credibility of alternate assessment results is required, including documentation of the reliability and validity of a state's alternate assessments, and increased rigor of standard setting processes for alternate assessment.
5. It is essential to consider how a state's approach reflects the assumption of standards-based reform that all children can learn, and schools can be held accountable for their learning.

   I think, in our school, for the first time, these students are seen as who they really are, individuals with a unique personality. This happened as soon as more of the staff and community became involved with them through standards-based instruction and alternate assessment. Standards and alternate assessments bring together the best skills of both general and special educators. Parents really love the collections of student work. It showcases a student's performance and helps us all see the growth that's really happening.

   (Thompson, Quenemoen, Thurlow, & Ysseldyke, 2001, composite of teacher comments from surveys after first year of implementation).

References

AERA/APA/NCME. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association, American Psychological Association, National Council on Measurement in Education.

Browder, D. (2001). *Curriculum and assessment for students with moderate and severe disabilities*. New York: Guilford Press.

Brualdi, A. (1999). *Traditional and modern concepts of validity*. Washington, DC: Retrieved May 20, 2002 from the ERIC/AE Digest on the World Wide Web: http://www.ed.gov/databases/ERIC_Digests/ed435714.html.

Cizek, G. (Ed.). (2001). *Setting performance standards: Concepts, methods, and perspectives*. Mahwah, NJ: Lawrence Erlbaum Associates.

Hill, R. (2001). The impact of including special education students in accountability systems. National Center for the Improvement of Educational Assessment, Inc. http://www.nciea.org/publications/CCSSOSpecialEd_Hill01.pdf

Heubert, J. P., & Hauser, R. M. (1999). *High stakes: Testing for tracking, promotion, and graduation.* Washington, DC: National Academy Press.

Kamfer, S., Horvath, L., Kleinert, H., & Kearns, J. (2001). Teachers' perceptions of one state's alternate portfolio assessment systems: Implications for practice and teacher preparation. *Exceptional Children, 67*(3), 361-375.

Kentucky state accountability system information taken from: http://www.lrc.state.ky.us/kar/703/005/020.htm, http://www.ihdi.uky.edu/kap/ , http://www.ihdi.uky.edu/kap/history.htm, and http://www.ihdi.uky.edu/kap/faq.asp#What%20is%20measured%20by%20the% 20Alternate%20Assessment

Kleinert, H., & Kearns, J. (2001). *Alternate assessment: Measuring outcomes and supports for students with disabilities*. Baltimore: Brookes Publishing.

Kleinert, H., & Kearns, J. (1999). A validation study of the performance indicators and learner outcomes of Kentucky's alternate assessment for students with significant disabilities. *Journal of The Association for Persons with Severe Handicaps, 24*(2), 100-110.

Kleinert, H., Kennedy, S., & Kearns, J. (1999). Impact of alternate assessments: A statewide teacher survey. *Journal of Special Education, 33*(2), 93-102.

Marion, S. (2001). *Wyoming's approach for including all students in the assessment and accountability system*. PowerPoint presentation at CCSSO Large-scale assessment conference, June 26, 2001.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13-103). New York: Macmillan.

Mid-Continent Educational Laboratory. (2000). *Content knowledge* (3rd ed.). Aurora, CO: McREL. Retrieved May 20, 2002, from the World Wide Web: http://www.mcrel.org/standards-benchmarks/docs/reference.asp

North Carolina state accountability system information taken from: http://www.ncpublicschools.org/abcs/ABCsHist.html#aa and http://www.ncpublicschools.org/accountability/

Quenemoen, R., Massanari, C., Thompson, S., & Thurlow, M. (2000). *Alternate assessment forum: Connecting into a whole.* Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

Roeber, E. (2002). *Setting standards on alternate assessments* (Synthesis Report 42). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

Thompson, S.J., Erickson, R., Thurlow, M.L., Ysseldyke, J., & Callender, S. (1999). *Status of the states in the development of alternate assessments* (Synthesis Report 31). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

Thompson, S.J., Quenemoen, R., Thurlow, M.L., & Ysseldyke, J.E. (2001). *Alternate assessments for students with disabilities*. Thousand Oaks, CA: Corwin Press.

Thompson, S.J., & Thurlow, M.L. (1999). *1999 State special education outcomes: A report on state activities at the end of the century*. Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

Thompson, S.J., & Thurlow, M.L. (2000). *State alternate assessments: Status as IDEA alternate assessment requirements take effect* (Synthesis Report 35). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

Thompson, S.J., & Thurlow, M.L. (2001). *2001 State special education outcomes: A report on state activities at the beginning of a new decade*. Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

Turner, M., Baldwin, L., Kleinert, H., & Kearns, J. (2000). An examination of the concurrent validity of Kentucky's alternate assessment system. *Journal of Special Education, 34*(2), 69-76.

Appendix

Side-by-Side Comparison of Literature Addressing the Development of Alternate Assessments

Browder, Diane. (2001). Curriculum and assessment for students with moderate and severe disabilities. New York: Guilford Press.

Kleinert, H. & Kearns, J. (2001). Alternate Assessment: Measuring Outcomes and Supports for Students with Disabilities. Baltimore: Brookes Publishing.

Quenemoen, R., Thompson, S., Thurlow, M., & Lehr, C. (2001). A Self-Study Guide to Implementation of Inclusive Assessment and Accountability Systems. Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes. http://education.umn.edu/nceo/OnlinePubs/workbook.pdf

Thompson, S.J., Quenemoen, R., Thurlow, M.L., & Ysseldyke, J.E. (2001). Alternate assessments for students with disabilities. Thousand Oaks, CA: Corwin Press.

Thurlow, M., Quenemoen, R., Thompson, S., & Lehr, C. (2001). *Principles and characteristics of inclusive assessment and accountability systems*(Synthesis Report 40). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

Essential Alternate Assessment Development Processes	**Alternate Assessment for Students with Disabilities(Thompson, Quenemoen, Thurlow, & Ysseldyke, 2001)**	**Alternate Assessment: Measuring Outcomes and Supports for Students with Disabilities (Kleinert & Kearns, 2001)**	**Curriculum and Assessment for Students with Moderate and Severe Disabilities (Browder, 2001)**	**NCEO Principles of Inclusive Assessment and Accountability Systems (Thurlow, Quenemoen, Thompson, & Lehr, 2001); A Self-Study Guide to Implementation of Inclusive Assessment and Accountability Systems (Quenemoen, Thompson, Thurlow, & Lehr, 2001)**
Stakeholder and policymaker development of desired student outcomes	Chapter 1: Consider how all of the students in your school can work toward the same content standards, and how their progress can be measured. Then, consider how students with significant disabilities are working toward broad content standards, and how an alternate assessment can measure their progress. You may find that your students are not getting as many opportunities to learn to the	p. 11 "…the question of what to assess poses considerable challenges. The question cannot be divorced from the context of the state or district content standards that are the framework for the general curriculum and the regular assessment…broadly stated content standards – focused on the broad application of core content to "real-life" contexts – are clearly more suited to inclusion in the alternate assessment than are more narrowly written standards that focus on only a prescribed set of	p. 3-4 "This book describes specific methods for conducting alternate assessment of students with disabilities that will meet state standards, developing an IEP, and assessing progress on target objectives to improve student learning. The approach…is consistent with trends in curriculum-based assessment described for students with mild disabilities…[but] for students with severe disabilities, professionals first must define	Alternate Assessment 1.a. Alternate assessments are aligned with state standards held for all students, through some process of extension, expansion, access, or other high performance bridge to the state content standards.

	standards as they should.	academic content." p. 47 "In special education there is a longstanding debate about what students with disabilities should and do know. We believe strongly that the general curriculum framework should be the first consideration for IEP teams.	an individualized curriculum before planning specific assessment. This specific assessment will often use behavioral assessment strategies, but may also include qualitative appraisals such as portfolio assessment."	
Careful development, testing, and refinement of assessment methods	Chapter 1: The nuts and bolts of gathering high quality assessment data, and assembling them in appropriate ways is the core of the process of alternate assessment, but gathering high quality data depends on the thoughtful preparatory work as the IEP is developed, as instruction is carried out, and as progress is observed and documented. States and districts vary on how data will be assembled and handled, so what you do specifically to gather and prepare the data will vary as well.	p. 12 "We believe very strongly that alternate assessments should be performance based ("testing methods that require students to create an answer or product that demonstrates their knowledge or skills," U.S. Congress, 1992), as opposed to more "paper-and-pencil" – based measures. . . Portfolio assessments, which are performance-based collections of student products, are especially suited for alternate assessments … [as one of several reasons] ...portfolio assessment enables students and teachers to use multiple measures…and can provide a broadly defined assessment structure capable of accommodating a very diverse student population." p. 13 "Alternate assessments should allow the student to apply what he or she has learned; skills are not used in isolation. Instead, they are parts of complex performances that integrate skills across developmental and academic areas…Alternate assessment should not be a one-time test or single snapshot of student performance."	p. 16-17, Table 1.3. Methods Used to Assess Students with Moderate and Severe Disabilities notes that "[alternate assessment] may use any of the above described assessment methods." An analysis of shortcomings and contributions to current practice of each approach is given for the multiple methods. The book focuses on a blended approach that is relevant both to student's life skills needs and to their public school experience.	Alternate Assessment 1.c. Alternate assessment options promote the use of a variety of valid authentic performance-based assessment strategies aligned to standards, allowing all students to be able to show what they know and are able to do.
Scoring of evidence according to professional accepted standards	Chapter 1: Once alternate assessment data are gathered, state, district, and school administrators must have all assessments scored, and then report to the public and to the parents how their school is doing. From chapter 7: These processes vary greatly, for example: (excerpts) Teachers are trained to score all the student work, but no one scores his or her own students or own district work, followed by expert rescoring of a sample of work; in some states, regional panels review all portfolios with at least two independent scores, and expert scorers rescore those portfolios whenever the original two scores are in disagreement.		p. 78 "First performance indicators should be clearly defined and validated with stakeholders…Second, clear guidelines should be developed for scoring… A third way to increase reliability and validity in portfolio assessment is to recruit evaluator (e.g., teachers) who know the types of programs being assessed, and to train them to reach high levels of agreement in using a checklist, rubric, or other scoring method. Whatever method is used to develop and score the alternate assessment, consideration needs to be given to how the results will be used." P. 80 "… alternate assessment information can be used for the benefit of students with disabilities if it becomes a foundation for quality enhancement for these educational programs."	Principle 3. All students with disabilities are included when student scores are publicly reported, in the same frequency and format as all other students, whether they participate with or without accommodations, or in an alternate assessment. This principle provides the first level of accountability for the scores of students with disabilities. Regardless of how students participate in assessments, with or without accommodations, or in an alternate assessment, students' scores are reported, or if scores are not reported due to technical issues or absence, the students are still accounted for in the reporting system.
Standard setting processes to allow use of results in	Chapter 7: States incorporate their "values" into rubric content			Principle 4. The assessment performance of students with disabilities has the same impact on the final accountability index as the performance of other

reporting and accountability systems	and scoring… After setting criteria, developing rubrics, and scoring the evidence, another step is taken to refine the "levels" of performance against the content and performance standards… states and districts work to develop a common understanding of "levels" of performance for students participating in alternate assessments. In Chapter 2 we explored the ideas that content and performance standards may be the same for all students; but performance indicators that demonstrate progress toward the standards will vary for students who have many challenges for learning… There are many variations to how levels are determined across the states, and the beliefs and philosophy of the state stakeholders typically drive these decisions.			students, regardless of how the students participate in the assessment system (i.e., with or without accommodations, or in an alternate assessment). This principle provides the second level of accountability for students with disabilities. In order for all students to count in increased expectations for accountable schools, all student assessment participation and performance data must be integrated into district and state accountability indices. Federal Title I requirements specifically require this, but districts and states should address fully inclusive accountability in any local or state developed accountability indices to promote equal access and opportunity for all students. Alternate Assessment 3.d. Scoring and reporting processes include a detailed approach for administration, with clearly defined performance standards, scoring and recording procedures, and reliability checks built into the process.
Continuous improvement of the assessment process	After completing an assessment year, and before moving into the next year, take time to check out the benefits and challenges, not only of the standards and assessment systems, but the benefits of the reform effort for students and everyone who serves them.	p. 225 "Performance-based assessment strategies for students with significant disabilities are hardly new…What is new, however, is the inclusion of students with significant disabilities in large-scale educational assessments, especially through the use of alternate assessments. We do not pretend to have the definitive answers as to how closely alternate assessments for these students can be tied to the learner outcomes identified for all students or what the performance and scoring criteria for these assessments should be. .. Research in each of these areas is in its infancy but fortunately these important questions are now being addressed for the good of all students now and in the future." See references section for citations on the Kentucky research on impact of alternate assessment	p. 80 The starting point in quality enhancement is to set the goal of having a high-quality program… Although there may be real problems with both the evaluation and the resources available, a team that is committed to excellence will respond to these challenges with problem solving, rather than viewing them as reasons not to respond. The fact that many states and districts with accountability systems did not include students with moderate and severe disabilities until the IDEA 1997 mandates suggests that these students were overlooked in initial discussions of school reform."	Principle 5. There is improvement of both the assessment system and the accountability system over time, through the processes of formal monitoring, ongoing evaluation, and systematic training in the context of emerging research and best practice. This principle addresses the need to base inclusive assessment and accountability practices on current and emerging research and best practice, with continuous improvement of practices as research-based understanding evolves. By working together on improvement of assessment and accountability systems, stakeholders can sustain commitment to keeping the standards high and keeping the focus clear on all students being successful. Ongoing training of IEP team members and other key partners is an essential component of this effort.